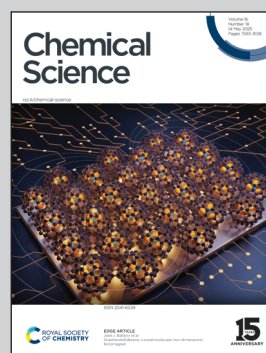Showcasing research from Professor Liu's laboratory, Department of Chemistry, Zhejiang University, Zhejiang, China.

Discovery of a molecular adsorbent for efficient CO_2/CH_4 separation using a computation-ready experimental database of porous molecular materials

A computation-ready database of 7,939 macrocycles and cages accelerates the discovery of porous molecular materials for CO_2/CH_4 separation, bridging structure, property, and function.

Image reproduced by permission of Ming Liu from *Chem. Sci.*, 2025, **16**, 7685.

ROYAL SOCIETY OF CHEMISTRY

15 YEARS ANNIVERSARY

rsc.li/chemical-science

Discovery of a molecular adsorbent for efficient CO₂/CH₄ separation using a computation-ready experimental database of porous molecular materials†

Siyuan Yang,‡[ab] Qianqian Mao,‡[a] Heng Ji,‡[a] Dingyue Hu,[a] Jinjin Zhang,[a] Linjiang Chen*[cd] and Ming Liu [iD] *[ab]

The development and sharing of computational databases for metal–organic frameworks (MOFs) and covalent organic frameworks (COFs) have significantly accelerated the exploration and application of these materials. Recently, molecular materials have emerged as a notable subclass of porous materials, characterized by their crystallinity, modularity, and processability. Among these, macrocycles and cages stand out as representative molecules. Experimental discovery of a target molecular material from a vast possibility of structures for defined applications is generally impractical due to high experimental costs. This study presents the most extensive Computation-ready Experimental (CoRE) database of macrocycles and cages (MCD) to date, comprising 7939 structures. Using the MCD, we conducted simulations of binary CO₂/CH₄ competitive adsorption under conditions relevant to industrial applications. These simulations established a structure–property–function relationship, enabling the identification of materials with potential for CO₂/CH₄ separation. Among them, a macrocycle, **NDI-Δ**, exhibited promising CO₂ adsorption capacity and selectivity, as confirmed by gas sorption and breakthrough experiments.

Introduction

Porous materials hold great promise to address challenges in achieving clean energy and sustainability. Experimentally exploring the target porous materials from a vast possibility of structures for defined applications is generally impractical due to high experimental costs. High-throughput computational screening has emerged as a valuable tool for discovering functionalities in these materials. Porous material databases play a pivotal role in catalyzing the rational design of porous materials for desired functions,[1–6] such as gas separation,[7–9] storage,[10–12] adsorption heat pumps[13] and so on. Wilmer and colleagues built a hypothetical MOF database (hMOF) of 137 953 MOFs constructed from 103 different building blocks.[14] The

database has been instrumental in numerous experimental discoveries of functional metal–organic frameworks (MOFs).[11,15] Colón *et al.* introduced the Topologically Based Crystal Constructor (ToBaCCo) database, featuring 13 512 MOFs with 41 unique topologies.[16] Concurrently, the development of COF databases has been remarkable, amassing over 560 000 structures.[17–20] Experimental MOF structures, like those found in the Cambridge Structural Database (CSD), often contain solvent molecules within their pores, exhibit positional disorders in certain fragments, and lack hydrogen atoms, among other issues. These factors make such MOF structures unsuitable for direct use in computational studies. To address this, Chung *et al.* created a database of Computation-ready Experimental (CoRE) MOFs.[21] In the CoRE-MOF database each MOF structure undergoes a multi-stage cleaning process to resolve structural issues, making them directly useable in calculations. The creation of the CoRE-MOF database featured systematic data curation and standardization procedures, and the approximately 14 000 structures[22] originating from experimental data are immediately ready for computational studies. The database subsequently proved useful for researchers seeking to explore the potential of MOFs in various applications through computational studies.[23–25] Later, Rosen *et al.* developed the Quantum MOF (QMOF) database, which consists of Density Functional Theory (DFT) optimized experimental MOFs

[a]*Department of Chemistry, Zhejiang University, Hangzhou, Zhejiang 310027, China. E-mail: mingliu@zju.edu.cn*

[b]*Hangzhou Global Scientific and Technological Innovation Center (HIC), Zhejiang University, Hangzhou, Zhejiang 311200, China*

[c]*Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China. E-mail: linjiangchen@ustc.edu.cn*

[d]*School of Chemistry and School of Computer Science, University of Birmingham, Birmingham, B15 2TT, UK. E-mail: linjiangchen@ustc.edu.cn*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5sc01532d

‡ These authors contributed equally to this work.

with a range of DFT derived properties for each MOF, such as band gaps, Density-Derived Electrostatic and Chemical protocol 6 (DDEC6) charges, spin densities, and more.[26]

Similarly, Tong *et al.* compiled a CoRE-COF database of now over 600 covalent organic frameworks (COFs), extracted from the experimental literature.[27] Based on CoRE structures, Ongari *et al.* developed a database comprising 324 "Clean, Uniform, and Refined with Automatic Tracking from Experimental Database" (CURATED) COF structures and updated it to 871 structures.[28] The CURATED COF database has further undergone an optimization process for both atomic coordinates and cell dimensions of the CoRE structures, employing a multi-step DFT approach. Subsequently, DDEC6 charges were assigned to each atom, enhancing the database's utility.

Apart from MOFs and COFs, molecular materials have recently emerged as a notable subclass of porous materials, characterized by their modularity and processability.[29] Unlike MOFs and COFs, which are extended framework structures, molecular materials are formed by the assembly of discrete molecules. The discrete feature provides the molecular materials with good solubility in common solvents or high solution dispersibility, and therefore promotes processability during applications, compared to MOFs and COFs.[30] As many of these molecules have been intensively investigated as hosts in supramolecular systems, there are a great number of single crystal structures that have been reported and archived, leaving a valuable resource for potential high throughput structure screening. Therefore, there are continuous efforts to develop new porous molecular materials *via* computational design and computational screening. Evans *et al.* suggested that using small organic molecules exclusively as the building blocks for cage-based porous molecular structures could yield up to 10^{60} possible variants.[31] This highlights the vast potential of using these organic entities for innovative porous molecular crystal discovery. Msayib *et al.* carried out a focused exploration within the CSD for molecular crystals with the capability for adsorption of hydrogen and nitrogen and identified 23 promising candidates.[32] There are several molecular crystal databases that have been developed, such as the organic porous molecular crystals database (oPMC)[33] and the Cage Database (CDB).[34] Recently, Li *et al.* established the first database of metal–organic cages (MOCs), containing 1839 structures, and also the largest database to date of experimental organic cages (OCs), containing 7736 cages.[35] This was achieved by integrating topological data analysis (TDA) and supervised and unsupervised learning methods. However, none of these databases entirely align with the criteria for ideal CoRE structures. Even for the most comprehensive OC database, a significant limitation arises from the presence of solvent molecules within many OCs, which obstruct channels or pores, thereby complicating the assessment of their potential applications. Moreover, the issue of redundant coordinates in the structure files, attributable to the occupancy ratio of atoms, requires meticulous correction and alignment with reference literature. Additionally, we noticed the inclusion of structures in the OC database that do not conform to the definitions of cages or macrocycles, necessitating their exclusion. Furthermore, the database's reliance on TDA to primarily consider the heaviest molecule within a structure has led to the unintended inclusion of rotaxanes and pseudorotaxanes.

This study introduces a CoRE database of macrocycles and cages, two of the most representative porous molecular materials. Using the CSD, we updated structures not previously catalogued within the OC. All structures were carefully curated and optimized in two steps, applying semi-empirical DFT to both atomic coordinates and cell dimensions. Subsequently, DFT-derived DDEC6 partial charges were assigned to each atom. This CoRE macrocycle and cage database (MCD) can be directly used for screening molecular systems for target functions. We conducted competitive Grand Canonical Monte Carlo (GCMC) simulations on a selected dataset from MCD to assess the selective adsorption efficiency of these structures for CO_2 over CH_4. Among the selected candidates, a macrocycle **NDI-Δ** was identified for its promising CO_2 adsorption capacity and selectivity, as confirmed by gas sorption and breakthrough experiments. These results demonstrate the effectiveness of the MCD database for identifying promising molecular candidates for a target application.

Results and discussion

Database construction

Fig. 1 shows the structured methodology used to construct the MCD, organized into four distinct phases: candidate collection, where potential structures are initially gathered; manual selection, involving the identification and selection of structures specifically containing either a cage or a macrocycle; structure cleaning, where solvent molecules and redundant atomic coordinates are removed to refine the structures; and structure optimization and DFT-derived partial charge assignment, where
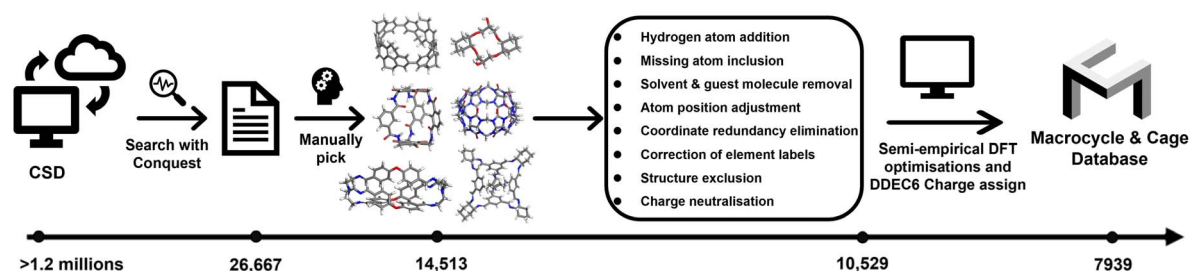

Fig. 1 Schematic illustration of the macrocycle and cage database construction.

the structures undergo optimization using semi-empirical DFT to optimize atomic coordinates and cell dimensions, followed by the assignment of DFT-derived partial charges to each atom.

Utilizing ConQuest (version 5.44, updated as of April 2023), we initiated a systematic search for potential candidates of cages and macrocycles, employing the 'must-have' criteria that were previously utilized for the OC database. The selection criteria were specific, requiring structures to have well-defined coordinates, to be non-polymeric and entirely organic, and to exclude any structures containing metal atoms. This selection process resulted in identifying 26 667 potential candidates, marking a notable increase from the 18 294 OC database candidates discovered using a previous version of the CSD. Details on the specific 'must-have' fragments and the corresponding numbers of hits are provided in Table S1.†

From the 26 667 candidates identified, 7736 had already been classified as 'organic cages' in the OC database, leaving 18 931 candidates with classifications yet to be determined. As previously discussed, the TDA method might categorize some structures that do not conform to the traditional definitions of cages or macrocycles. Furthermore, there is a deliberate effort to exclude rotaxanes and pseudorotaxanes, given the reported limited porosity of some macrocycles. Considering these factors, we undertook a manual review of the 18 931 unclassified structures. This review process aimed to identify and select those candidates that feature at least one macrocycle or cage structure, ensuring the relevance and accuracy of our database's content.

To ensure that structures from the CSD are suitable for computational simulations, a comprehensive cleaning and correction process was implemented to achieve a 'computationally ready' status. This process included the addition of hydrogen atoms, completion of missing atoms, removal of solvents and small guest molecules, adjustment of atomic positions, elimination of redundant atoms, correction of elemental assignments, structure exclusion and charge neutralization. These structural cleaning processes are detailed further in the Experimental section and ESI Section S2.†

After the structure cleaning process, more than 10 500 configurations were curated. However, the removal of solvent molecules may inadvertently increase the porosity of these configurations, which could lead to biased outcomes in simulations. Therefore, it is necessary to perform geometry optimization on these configurations to ensure accuracy. The application of DFT for optimization across such a large dataset presents considerable computational challenges, especially for structures generally comprising thousands of atoms. The GFNn-xTB series, particularly the GFN2-xTB model, has been shown to offer remarkable accuracy in geometry reproduction across a diverse array of systems compared to other semi-empirical methods.[36,37] Its effectiveness has been demonstrated in the geometry optimization of large structures, such as transition metal complexes,[38] periodic peptides and proteins,[39] proving its capability to accurately reproduce structures as confirmed by X-ray diffraction data. GFN2-xTB was used to optimize the curated structures in two steps. First, the atomic coordinates were

optimized with fixed cell parameters, followed by a second optimization process allowing full structural flexibility.

Notably, structural errors – whether missing elements or inaccuracies introduced during the cleaning phase – were identified through warnings in the optimization convergence process. These issues were then rectified, and the affected structures were reprocessed.

Finally, DFT-derived DDEC6 partial charges were computed for the optimized structures, utilizing the electron density that was computed by Perdew–Burke–Ernzerhof (PBE)[40]/DZVP-MOLOPT-PBE-GTH basis sets[41] with DFT-D3(BJ) dispersion corrections.[42]

Overview of MCD

Cages, regarded as a distinct type of macrocycle, have garnered considerable interest in the area of supramolecular chemistry and recently as an emerging subset of porous materials. In this work, we explore cages as a separate entity and compare their characteristics with other types of macrocycles. Our criterion for classifying cages and macrocycles is defined as follows: if a macrocycle contains three or more windows that share the same void space, as illustrated in Fig. 2a, it is considered a cage. Otherwise, it is classified as a macrocycle (Fig. 2b). MCD contains a total of 7939 cleaned and optimized molecular crystal structures, including 6679 macrocycles and 1260 cages. The elemental composition of the structures catalogued in the MCD (Fig. 2c) highlights a diverse range of organic elements. The structures are predominantly composed of carbon, hydrogen, oxygen and nitrogen – the elements most commonly associated with organic molecules. The majority of the structures, 7898 out of 7,939, have fewer than 1000 atoms per unit cell (Fig. 2d). FT-RCC3,[43] an amine cage with the identifier VOMPAQ, has the highest number of atoms per unit cell in this database, containing 1584 atoms. This can be a useful metric for computational chemists when preparing resources for computational screening.

MCD includes both single-molecule crystals and cocrystals, with approximately 25% (1965 out of 7939) of the structures in the MCD containing more than one kind of molecule, highlighting
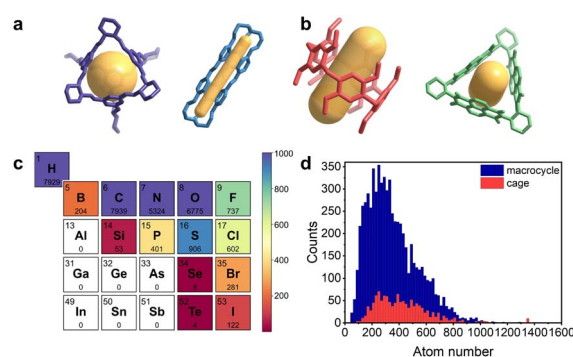

Fig. 2 (a) Examples of cages in MCD. (b) Examples of macrocycles in MCD. The voids within cages and macrocycles are highlighted in yellow. (c) The number of macrocycles and cages in the MCD containing specific elements. (d) Distribution of atomic numbers per unit cell across the structures in the MCD.

the comprehensive coverage and versatility of the database in capturing the structural diversity of macrocycle and cage crystals.

Validation of the optimization method

To assess the performance of the GFN2-xTB optimization method, 800 structures previously optimized using GFN2-xTB were subjected to further optimization *via* DFT. Fig. 3 provides a detailed comparative analysis between the GFN2-xTB and DFT optimization results. Of these structures, 767 achieved convergences in the DFT optimization process. Energy differences per atom between the two methods are small (Fig. 3a), with the total energy changes being less than 0.01 eV in 90% of cases, and over 98% of the structures showing dispersion energy changes of less than 0.005 eV. The 'Superpose Structure' technique in Materials Studio[44] facilitated the examination of structural similarities by overlaying $2 \times 2 \times 2$ supercell structures and assessing the similarity for each structure pair. To clearly represent the similarity comparison between GFN2-xTB and DFT optimized structures, the similarities were classified into four stages in descending order: stage I (90–100%), stage II (80–90%), stage III (70–80%), and stage IV (less than 70%). As shown in Fig. 3b, the majority of structures optimized by GFN2-xTB closely match those optimized by DFT, with more than 88% of the comparisons showing a similarity greater than 97%. Conversely, only about 2.7% of the structures show a similarity below 90%, indicating a high degree of consistency between the two optimization methods.

The changes of cell volumes, the diameter of the largest inclusion sphere (D_i), and the pore limiting diameter (D_f) between structures optimized by GFN2-xTB and those further optimized by DFT are insignificant (Fig. 3c to e). The correlation coefficients (R^2) for the linear relationship $y = x$ for cell volume, D_i, and D_f are exceptionally high, standing at 0.99 for cell volume and 0.98 for both D_i and D_f. This indicates a strong agreement between the two sets. Specifically, for cages, the consistency across all three geometric parameters is remarkable, with an R^2 value of 0.99, denoting very slight variations in most cases. However, the structure REWKIQ demonstrates a notable deviation, particularly in D_i, where it underwent a significant contraction from 4.33 Å to 2.19 Å (nearly a 50% reduction) after DFT optimization. Additionally, D_f saw a decrease from 1.38 Å to 0.99 Å. As a result, the structural similarity between the DFT and GFN2-xTB-optimized configurations of REWKIQ is relatively low, at 77.28%. This particular case of REWKIQ also highlighted the most substantial change in dispersion energy, at -0.0101 eV per atom, with a total energy difference of -0.0130 eV. These energy variations suggest that the discrepancies observed are primarily due to changes in molecular movement and denser packing after DFT optimization, which significantly affect the dimensions of D_i and D_f. The details of REWKIQ structural comparison can be seen in Fig. S4.†

The accessible volume (Fig. 3f) and surface area (Fig. 3g) of GFN2-xTB optimized structures underwent slightly larger adjustments during DFT optimization than cell volume, D_i and D_f. The R^2 values for accessible volume and surface area are 0.97 and 0.95, respectively, indicating a strong linear relationship despite these adjustments. Notably, for cages, the R^2 values remain exceptionally high at 0.99. This high level of consistency may be attributed to the predefined pore and window structures
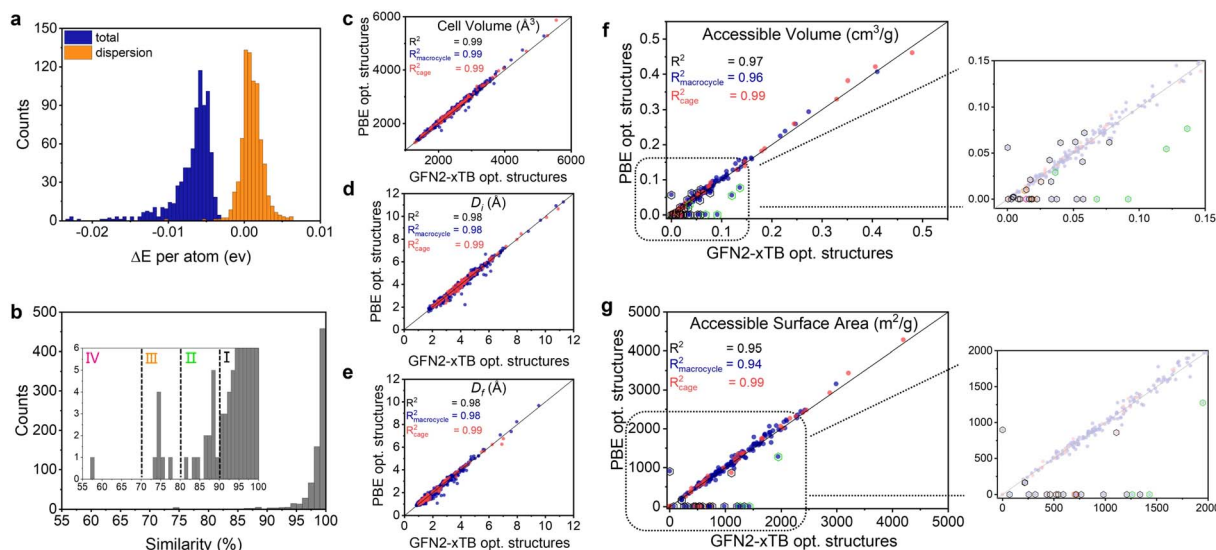

Fig. 3 Comparison between structures optimized using GFN2-xTB and those further refined through DFT optimization. (a) Histogram illustrating the distribution of energy differences between the structures before and after DFT optimization. (b) Assessment of the structural similarity between the optimized configurations. Comparisons regarding cell volume (c), D_i (d), D_f (e), accessible volume (f), and accessible surface area (g). Macrocycles are indicated by navy solid circles, whereas cages are denoted by red solid circles. To evaluate the accessibility of pore volume and surface area, a probe with a radius of 1.65 Å, corresponding to the kinetic radius of CO_2, was employed. Structures exhibiting more than a 20% change in accessible volume or surface area are highlighted with open hexagons. The colour coding of these hexagons—black for stage I (90% ≤ similarity < 100%), green for stage II (80% ≤ similarity < 90%), orange for stage III (70% ≤ similarity < 80%), and magenta for stage IV (similarity <70%).

characteristic of cage molecules. Interestingly, after DFT optimization, 17 structures were identified as non-porous, of which 13 belonged to similarity stage I, indicating that their similarity to their GFN2-xTB-optimized counterparts exceeded 90%. Within this subset, FAFHUQ exhibited the lowest similarity at 95%. On the other hand, GUNDEZ underwent a transition from non-porous to porous as a result of DFT optimization, achieving a similarity of 94% alongside an accessible volume of 0.056 cm^3 g^{-1} and a surface area of 898.3 m^2 g^{-1}. These observations underline the potential for even minor modifications in molecular crystal structures to significantly impact their geometric properties. Such changes are crucial considerations in the preliminary screening process for material selection, demonstrating the importance of recognizing the flexibility and dynamic properties inherent in molecular crystals.

Impact of optimization on cleaned structures

Geometric dimensions generally tend to decrease after optimization. Specifically, cell volume (Fig. 4a) was reduced in approximately 89% of the structures (7082 out of 7939). Similarly, the D_i (Fig. 4b) and D_f (Fig. 4c) saw reductions in around 74% of cases (5879 and 5880 out of 7939, respectively). This trend towards smaller dimensions primarily stems from the removal of solvent molecules during the cleaning process, which was conducted on 4909 structures. Eliminating these molecules left voids that were subsequently minimized by closer packing of the host molecules during the optimization step, resulting in reduced cell volumes in 91% of cases where solvents were removed. This highlights the necessity of structural optimization following structural cleaning.

The changes in accessible surface area, volume, and void fraction are shown in Fig. 4d to f. From the dataset, 1789 structures were initially identified as being porous to CO_2 in their cleaned state, but this number fell to 1142 after optimization. 712 structures that were initially porous lost their porosity after optimization, while 65 structures changed from being non-porous to porous. Reflecting the patterns seen in cell volume, D_i, and D_f, a significant proportion of structures experienced reductions in accessible surface area (90%), accessible volume (82%), and void fraction (80%). This analysis explicitly excludes structures categorized as non-porous both before and after optimization.

High-throughput screening of CO_2-selective adsorbents and experimental validation

Natural gas sweetening, the separation of CO_2 from CO_2/CH_4 mixtures, is recognized as a promising method to mitigate anthropogenic CO_2 emissions.[45] However, the size difference between the two gas molecules is relatively small, at only 0.5 Å, making it challenging to identify or design a porous structure suitable for CO_2/CH_4 separation. Experimentally investigating materials in MCD, including synthesis, crystallization, and adsorption isotherm measurements, is both time-consuming and labour-intensive.

The efficiency of computationally driven material discovery using high-throughput GCMC simulations based on other CoRE databases has been demonstrated in several previous studies.[46,47] First, we selected structures in MCD that have been previously studied for CO_2 and CH_4 adsorption isotherms in the literature, and the simulated isotherms for both CO_2 and CH_4
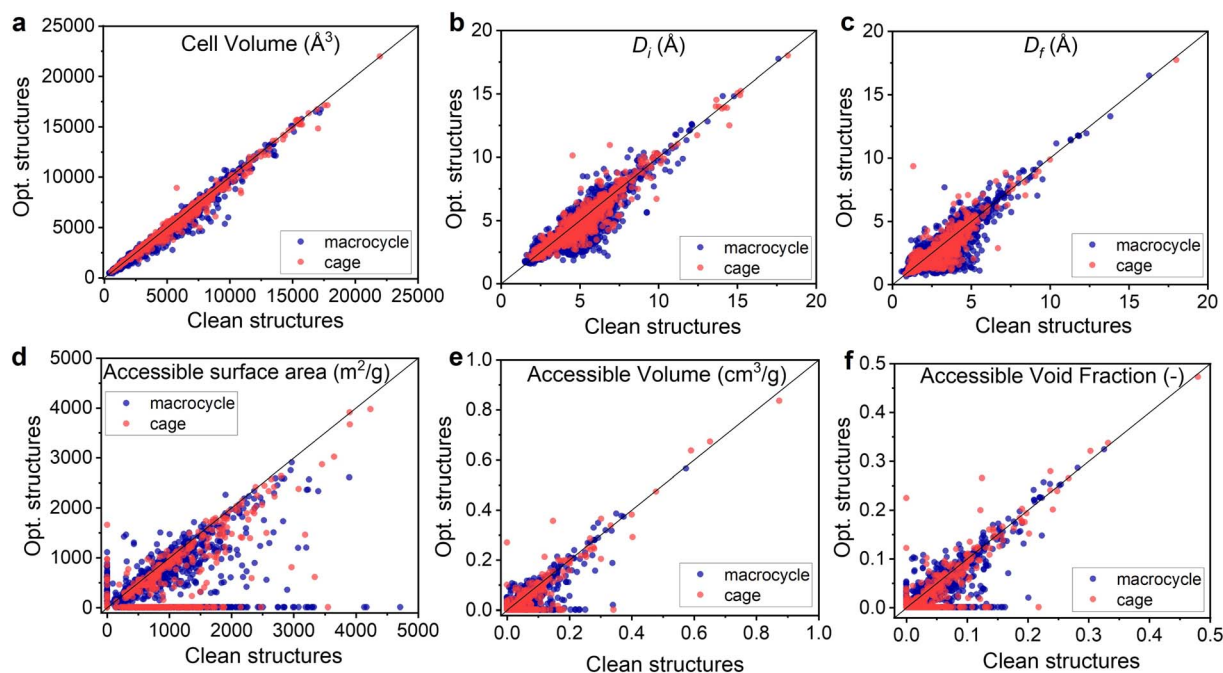
Fig. 4 Parity plots that contrast the geometric properties of macrocycles (in navy) and cages (in red) before and after optimization. These plots encompass various properties including cell volume (a), D_i (b), D_f (c), accessible surface area (d), accessible volume (e), and accessible void fraction (f). To evaluate the pore volume and surface accessibility, a spherical probe with a radius of 1.65 Å, corresponding to the kinetic radius of CO_2, was utilized.

showed good agreement in shape with the experimental adsorption isotherms (Table S4†). The adsorption mechanisms of these materials were captured by the GCMC simulations. The IAST selectivity values from simulated and experimental isotherms are comparable, as is the selectivity ranking. However, due to the inherent flexibility[48–51] of molecular materials and the loss of crystallinity after activation, GCMC simulations are not able to precisely reproduce the exact adsorption quantities of all materials. The selectivity and ranking derived from simulations remain valid for guiding the identification of promising materials. We conducted high-throughput GCMC simulations to assess the competitive adsorption of CO_2/CH_4, aiming to demonstrate that MCD is capable of identifying potential materials for specific applications. Criteria for selection included a requirement for structures in MCD to have a D_f exceeding 3.80 Å,[52] the kinetic diameter of CH_4, leading to a subset of 697 structures for analysis.

Fig. 5a visualizes the correlation between CO_2 uptake and CO_2/CH_4 selectivity across these structures, revealing that materials exhibiting the highest selectivity generally possess smaller D_f values. The materials displaying the top five selectivity were identified as DORZAO (4.25 Å), HOWNEO (5.62 Å), PUNMUH (5.47 Å), CUVHOT (4.98 Å), and OQIVAO (6.19 Å), arranged in descending order of selectivity. Notably, a larger accessible surface area was not indicative of increased CO_2 uptake under the conditions tested. The structures with the highest CO_2 uptakes had surface areas of 1181 m² g⁻¹

(REDMET), 1456 m² g⁻¹ (CUMHUO), 1455 m² g⁻¹ (KODMIC), 1097 m² g⁻¹ (WASPEO), and 2581 m² g⁻¹ (FINCIP), respectively. More detailed structure–property relationships can be found in ESI Section S7.†

DORZAO, the γ polymorph of (−)-NDI-Δ, is a macrocycle first reported by Stoddart's group in 2013.[53] For simplicity, we refer to it as NDI-Δ throughout the text. It stood out as the structure with the highest selectivity within the screening range, attaining a selectivity value of 50.14 and a CO_2 adsorption capacity of 1.97 mmol g⁻¹, with negligible CH_4 adsorption. Fig. 5b and c highlight two types of adsorption sites: intrinsic and extrinsic pores, in the simulated crystal structure after CO_2/CH_4 competitive adsorption. Notably, CH_4 adsorption within the intrinsic pore was virtually absent, as observed in the simulation movie outputs from RASPA. This result indicates CO_2's energetic preference for adsorption within this triangular pore structure, illustrating the structure's specificity and effectiveness for selective CO_2 adsorption.

According to the synthetic method reported in the literature,[53,54] we successfully prepared the rigid triangular macrocycle NDI-Δ by reacting naphthalenediimides (NDIs) with (RR)-trans-1,2-cyclohexanediamine. However, we were unable to obtain the γ-NDI-Δ polymorph under the same crystallization conditions after many attempts. Instead, a new phase, which we named γ′-NDI-Δ, with reasonably good crystallinity was obtained (Fig. S6†). CO_2 and CH_4 adsorption/desorption isotherms revealed relatively low adsorption capacities of 0.31
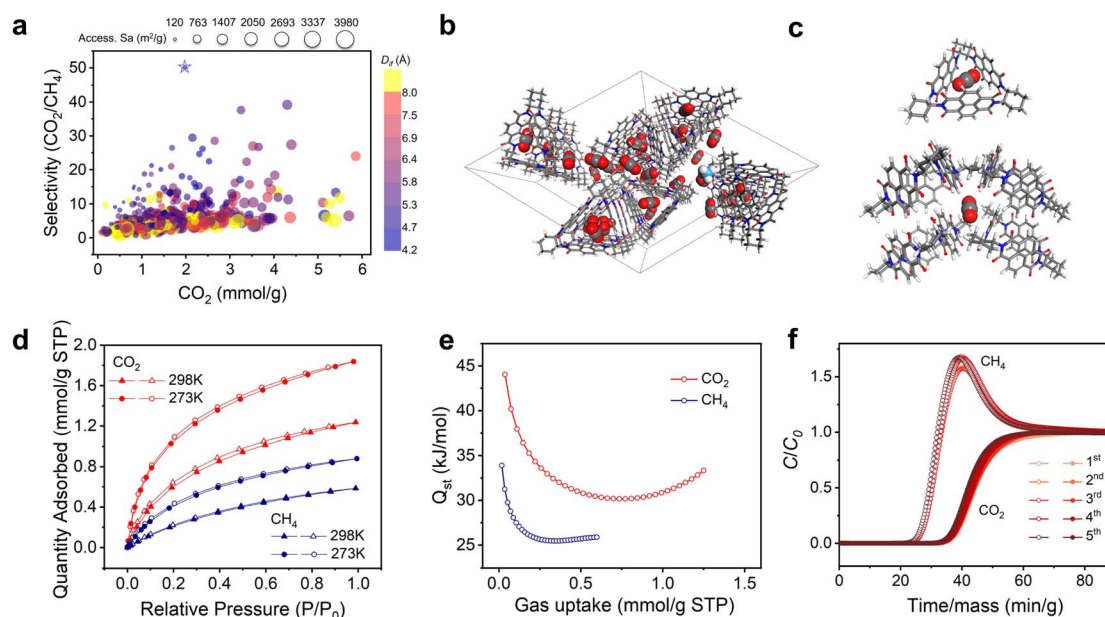

Fig. 5 (a) A coloured bubble map illustrating the relationship between CO_2 adsorption capacities and selectivity. The size of each bubble corresponds to the accessible surface area of the structure, while the colour gradient, ranging from blue to pink, represents the diameter of the largest inclusion sphere among the free path (D_{if}) values, with a maximum value capped at 8.0 Å. Structures with a D_{if} greater than 8.0 Å are highlighted in yellow. (b) An adsorption snapshot of molecules within the DORZAO structure, (c) showing both intrinsic (top) and extrinsic (bottom) adsorption sites. The structural representation uses a stick model for the host molecule and a CPK model for the guest molecules, with carbon in grey, oxygen in red, nitrogen in blue, hydrogen in white, and methane's carbon in cyan. (d) CO_2 and CH_4 sorption isotherms for NDI-Δ-CH_2Cl_2 (closed symbols: adsorption; open symbols: desorption) at 273 K and 298 K. (e) Isosteric heats of CO_2 and CH_4 adsorption for NDI-Δ-CH_2Cl_2, calculated using the Clausius–Clapeyron equation. (f) Separation performance of NDI-Δ-CH_2Cl_2: breakthrough curves for an equimolar binary mixture of CO_2/CH_4, with a flow rate of 1 mL min⁻¹ at 1 bar and 298 K.

mmol g^{-1} and 0.08 mmol g^{-1}, respectively (Fig. S7†). This lower capacity may be attributed to differences in molecular packing between γ′-**NDI-Δ** and γ-**NDI-Δ**, where the imperfect packing in γ′-**NDI-Δ** likely hinders access to adsorption sites as expected in γ-**NDI-Δ** from the database. Despite the low adsorption capacity, the breakthrough experiments of γ′-**NDI-Δ** at 1 bar and 298 K demonstrated promising CO$_2$/CH$_4$ selectivity. The breakthrough times were 14 min g^{-1} for CH$_4$ and 20 min g^{-1} for CO$_2$ (Fig. S8†).

Further efforts to obtain additional **NDI-Δ** phases were made. A second phase, named **NDI-Δ-Evap** (Fig. S6†), was prepared by evaporating the corresponding solution in CH$_2$Cl$_2$. This phase exhibited slightly improved gas adsorption performance compared to γ′-**NDI-Δ** (Fig. S7†). A third phase, **NDI-Δ-CH$_2$Cl$_2$**, previously reported in the literature,[54] was obtained by slow diffusion of CH$_3$OH into a CH$_2$Cl$_2$ solution of **NDI-Δ**. The PXRD patterns of the synthesized **NDI-Δ-CH$_2$Cl$_2$** matched well with simulated results from single-crystal data, confirming its phase purity (Fig. S9†). Gas adsorption studies showed that **NDI-Δ-CH$_2$Cl$_2$** displayed decent CO$_2$ adsorption capacities of 1.84 mmol g^{-1} (273 K) and 1.24 mmol g^{-1} (298 K) (Fig. 5d), despite becoming amorphous after activation.

The isosteric heat of adsorption (Q_{st}) at zero coverage was 44.05 kJ mol^{-1} for CO$_2$ and 33.89 kJ mol^{-1} for CH$_4$, indicating the inherent selectivity of **NDI-Δ** toward CO$_2$. More importantly, dynamic breakthrough experiments with a CO$_2$/CH$_4$ (1 : 1, v/v) mixture at 1 bar and 298 K (Fig. 5f) showed a 14 min g^{-1} breakthrough interval between CH$_4$ (20 min g^{-1}) and CO$_2$ (34 min g^{-1}), suggesting that the screened molecular adsorbent **NDI-Δ** can efficiently separate CO$_2$ from CH$_4$ at one breakthrough cycle. Additionally, the breakthrough curves remained stable over at least five cycles without significant changes in retention times, demonstrating the material's good reusability.

Conclusion

We have established a comprehensive workflow to build a database specifically for molecular crystals, with a focus on macrocycles and cages. The database contains structures that have been manually verified and optimized. The optimization process involved refining atomic orientations and cell dimensions using a two-stage semi-empirical DFT method, which demonstrated strong agreement with structures optimized using full DFT approaches in terms of both energy and geometry. Additionally, DDEC6 charges were assigned to each atom within these structures. With 7939 structures, the database is currently the most comprehensive collection of porous molecular materials accessible to the research community. We are dedicated to continuously enhancing and updating this resource in alignment with new developments in the CSD, ensuring its ongoing relevance and utility for macrocycle and cage exploration and discovery.

More importantly, we are able to perform high-throughput GCMC simulations based on the database to screen out target molecules for specific applications, such as competitive adsorption of CO$_2$ and CH$_4$. From the high-throughput

computational screening of a dataset selected from the database, we can readily identify a macrocycle, **NDI-Δ**, which exhibited promising CO$_2$ adsorption capacity and selectivity toward CO$_2$ over CH$_4$ even in its amorphous phase. It is worth noting that it would be very difficult to predict that such a macrocycle could efficiently separate CO$_2$ from CH$_4$ based solely on its chemical structure, even with the extensive knowledge of its rich host–guest chemistry. This demonstration underscores the database's potential to accelerate the discovery of functional molecular materials.

Methods

Structure curation

To ensure that structures from the CSD are suitable for computational simulations, a comprehensive cleaning and correction process was implemented to achieve a 'computationally ready' status. The curation processes undertaken included:

(1) Hydrogen atom addition: hydrogen atoms were added to structures missing hydrogen coordinates using BIOVIA Materials Studio,[44] to complete the molecular framework.

(2) Missing atom inclusion: absent atoms were added to the structures to maintain structural integrity.

(3) Solvent and guest molecule removal: solvent and small guest molecules were extracted, except in instances where an equivalent clean macrocycle or cage with similar packing was already documented in the MCD.

(4) Atom position adjustment: atom positions were altered to correct elongated, unrealistic bonds, thereby preventing errors in subsequent optimization steps.

(5) Coordinate redundancy elimination: redundant coordinates within structures have been removed.

(6) Correction of element labels: wrong element labels within the structures have been corrected.

(7) Structure exclusion: structures that were too messy to correct, did not contain macrocycles or cages, contained fullerenes, rotaxanes or those with coordinatively bonded metal elements were excluded.

(8) Charge neutralisation: the net charges of the structures were evaluated and adjusted to neutrality.

Two-step geometry optimization

The geometry optimization of the cleaned structures was executed in two steps using GFN2-xTB, as implemented in DFTB+ version 23.1.[55] Initially, the atomic coordinates of the structures were optimized with fixed cell parameters. This step was followed by a second optimization, allowing full structural flexibility. The optimization employed the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm, adhering to specific convergence criteria: an energy change (E_{conv}) threshold of 1×10^{-6} hartree and a residual force (G_{conv}) threshold of 1×10^{-4} hartree per bohr. Structures meeting these criteria advanced to subsequent phases. Optimizations were capped at 5000 steps and enforced a self-consistent charge (SCC) tolerance of 1×10^{-7} E_h, following DFTB+ guidelines,

with a maximum of 200 iterations. The settings for K-points were established using the "SupercellFolding" method in the DFTB+ algorithm, with further details provided in ESI Section S5.†

DDEC charge calculation

The computation of the DDEC6 (ref. 56 and 57) charge for each optimized structure was performed using Chargemol,[58,59] based on electron density profiles generated by DFT energy calculation using CP2K. The DFT calculations used the PBE exchange correlation functional,[40] extended by DFT-D3(BJ) dispersion corrections.[42] The calculations were facilitated by the Quickstep module within CP2K,[60] using Goedecker–Teter–Hutter (GTH) pseudopotentials[61] and the DZVP-MOLOPT-PBE-GTH basis sets.[41] A plane-wave cut-off of 600 Ry was set, along with a 4-level multigrid with relative cut-off of 50 Ry and a multiplication factor of 3. The Broyden diagonalization was used. The large scale CP2K input files were generated using Multiwfn 3.8 (dev) code.[62]

Validation of the geometry optimization

After completing two rounds of semi-empirical DFT geometry optimization, a subset of 800 structures – each containing between 150 and 300 atoms within its unit cell – was randomly selected for DFT geometry optimization. Optimizations were carried out under cell dimension unconstrained conditions at an external pressure of 1 atm. The optimization was performed using the L-BFGS optimizer, with convergence determined by specific criteria: the optimization was considered complete when the maximum change in geometry was less than 3×10^{-3} bohr, the root mean square (RMS) deviation was less than 1.5×10^{-3} bohr, the maximum force exerted on the atoms was less than 4.5×10^{-4} hartree per bohr, and the RMS deviation was less than 3×10^{-4} hartree per bohr. The energy calculation criteria remained consistent with previous settings. The similarity between structures optimized using DFT and those processed by the GFN2-xTB method was assessed with the "Superpose Structures" tool in Materials Studio, employing the field method.[44] To account for differences in intermolecular distances within the similarity evaluations, this analysis was conducted on $2 \times 2 \times 2$ supercell structures. This approach ensured a comprehensive and precise comparison of the structural similarities resulting from the two optimization methods.

Geometric-based descriptors

The topological analysis of the structures was performed utilizing Zeo++,[63] which employs Voronoi decomposition to calculate the geometric parameters of each structure's pore space. This analysis determines three key measurements: D_i, D_f, and D_{if}. For these calculations, a probe radius of 1.65 Å (ref. 52) (kinetic diameter of CO_2) was chosen to assess the accessible volume, void fraction, and surface area. To account for the accessibility of the pores specifically to CO_2 and CH_4 molecules, inaccessible pockets within the structures were blocked, using probe radii of 1.65 Å for CO_2 and 1.90 Å for CH_4,[52] respectively.

High-throughput competitive CO_2/CH_4 adsorption screening

A selection of structures from the MCD with a D_f exceeding 3.80 Å – the kinetic diameter of CH_4 – were chosen. This criterion ensures the unhindered passage of both CO_2 and CH_4 molecules through these structures. High-throughput GCMC simulations were conducted using RASPA 2.0.47,[64] taking into account the intermolecular interactions through the application of 6–12 Lennard-Jones (LJ) potentials, with a cutoff distance set at 12 Å. The LJ parameters for the host atoms were sourced from the DREIDING force field,[65] while CO_2 (ref. 66) and CH_4 (ref. 67) molecules were modelled using the TraPPE force field parameters. The simulations operated at a $50:50$ $CO_2:CH_4$ molar ratio, under conditions of 1 bar and 298 K. In these simulations, the structures of the macrocycles and cages were treated as rigid bodies. The Lorentz–Berthelot mixing rules were applied for interaction potentials, and electrostatic interactions were accurately depicted using the coulomb potential and Ewald summation. Each simulation ran for 10 000 cycles for equilibration and 200 000 cycles for production, executing a range of movements including translation, rotation, regrowth, identity changes and swap moves.

Data availability

All the crystal information files with DDEC6 charge, geometrical properties, results for macrocycle and cage identification and curation operations performed on each material are available at: **https://github.com/siyuanyang11/MCD** and **https://www.mingliulab.com/MCD**.

Author contributions

S. Y.: data curation, formal analysis, investigation, methodology, visualization, and writing – original draft. Q. M.: methodology, visualization, and writing – original draft. H. J.: data curation. D. H.: methodology and visualization. J. Z.: visualization. L. C.: conceptualization, resources, project administration, supervision, and writing – review & editing. M. L.: conceptualization, funding acquisition, project administration, resources, supervision, and writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Notes and references

1 P. Canepa, C. A. Arter, E. M. Conwill, D. H. Johnson, B. A. Shoemaker, K. Z. Soliman and T. Thonhauser, *J. Mater. Chem. A*, 2013, **1**, 13597.

2 Y. J. Colón and R. Q. Snurr, *Chem. Soc. Rev.*, 2014, **43**, 5735–5749.

3 T. Yan, Y. Lan, M. Tong and C. Zhong, *ACS Sustain. Chem. Eng.*, 2019, **7**, 1220–1227.

4 A. N. V. Azar, S. Velioglu and S. Keskin, *ACS Sustain. Chem. Eng.*, 2019, **7**, 9525–9536.

5 P. G. Boyd, A. Chidambaram, E. García-Díez, C. P. Ireland, T. D. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S. M. Moosavi, M. M. Maroto-Valer, J. A. Reimer, J. A. R. Navarro, T. K. Woo, S. Garcia, K. C. Stylianou and B. Smit, *Nature*, 2019, **576**, 253–256.

6 O. F. Altundal, C. Altintas and S. Keskin, *J. Mater. Chem. A*, 2020, **8**, 14609–14623.

7 Z. Qiao, K. Zhang and J. Jiang, *J. Mater. Chem. A*, 2016, **4**, 2105–2114.

8 P. G. Boyd, A. Chidambaram, E. García-Díez, C. P. Ireland, T. D. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S. M. Moosavi, M. M. Maroto-Valer, J. A. Reimer, J. A. R. Navarro, T. K. Woo, S. Garcia, K. C. Stylianou and B. Smit, *Nature*, 2019, **576**, 253–256.

9 A. H. Farmahini, S. Krishnamurthy, D. Friedrich, S. Brandani and L. Sarkisov, *Chem. Rev.*, 2021, **121**, 10666–10741.

10 H. Zhang, P. Yang, D. Yu, K. Wang and Q. Yang, *Chin. J. Chem. Eng.*, 2021, **39**, 286–296.

11 N. S. Bobbitt, J. Chen and R. Q. Snurr, *J. Phys. Chem. C*, 2016, **120**, 27328–27341.

12 A. W. Thornton, C. M. Simon, J. Kim, O. Kwon, K. S. Deeg, K. Konstas, S. J. Pas, M. R. Hill, D. A. Winkler, M. Haranczyk and B. Smit, *Chem. Mater.*, 2017, **29**, 2844–2854.

13 W. Li, X. Xia and S. Li, *ACS Appl. Mater. Interfaces*, 2020, **12**, 3265–3273.

14 C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp and R. Q. Snurr, *Nat. Chem.*, 2012, **4**, 83–89.

15 C. E. Wilmer, O. K. Farha, Y.-S. Bae, J. T. Hupp and R. Q. Snurr, *Energy Environ. Sci.*, 2012, **5**, 9849.

16 Y. J. Colón, D. A. Gómez-Gualdrón and R. Q. Snurr, *Cryst. Growth Des.*, 2017, **17**, 5801–5810.

17 R. L. Martin, C. M. Simon, B. Smit and M. Haranczyk, *J. Am. Chem. Soc.*, 2014, **136**, 5006–5022.

18 R. L. Martin, C. M. Simon, B. Medasani, D. K. Britt, B. Smit and M. Haranczyk, *J. Phys. Chem. C*, 2014, **118**, 23790–23802.

19 R. Mercado, R.-S. Fu, A. V. Yakutovich, L. Talirz, M. Haranczyk and B. Smit, *Chem. Mater.*, 2018, **30**, 5069–5086.

20 Y. Lan, X. Han, M. Tong, H. Huang, Q. Yang, D. Liu, X. Zhao and C. Zhong, *Nat. Commun.*, 2018, **9**, 5274.

21 Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl and R. Q. Snurr, *Chem. Mater.*, 2014, **26**, 6185–6192.

22 Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, B. Slater, J. I. Siepmann, D. S. Sholl and R. Q. Snurr, *J. Chem. Eng. Data*, 2019, **64**, 5985–5998.

23 H. Demir, S. J. Stoneburner, W. Jeong, D. Ray, X. Zhang, O. K. Farha, C. J. Cramer, J. I. Siepmann and L. Gagliardi, *J. Phys. Chem. C*, 2019, **123**, 12935–12946.

24 Y. Pramudya, S. Bonakala, D. Antypov, P. M. Bhatt, A. Shkurenko, M. Eddaoudi, M. J. Rosseinsky and M. S. Dyer, *Phys. Chem. Chem. Phys.*, 2020, **22**, 23073–23082.

25 Y. Wang, Z.-J. Jiang, D.-R. Wang, W. Lu and D. Li, *J. Am. Chem. Soc.*, 2024, **146**, 6955–6961.

26 A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein and R. Q. Snurr, *Matter*, 2021, **4**, 1578–1597.

27 M. Tong, Y. Lan, Z. Qin and C. Zhong, *J. Phys. Chem. C*, 2018, **122**, 13009–13016.

28 D. Ongari, A. V. Yakutovich, L. Talirz and B. Smit, *ACS Cent. Sci.*, 2019, **5**, 1663–1675.

29 T. Hasell and A. I. Cooper, *Nat. Rev. Mater.*, 2016, **1**, 16053.

30 X. Yang, Z. Ullah, J. F. Stoddart and C. T. Yavuz, *Chem. Rev.*, 2023, **123**, 4602–4634.

31 J. D. Evans, K. E. Jelfs, G. M. Day and C. J. Doonan, *Chem. Soc. Rev.*, 2017, **46**, 3286–3301.

32 K. J. Msayib, D. Book, P. M. Budd, N. Chaukura, K. D. M. Harris, M. Helliwell, S. Tedds, A. Walton, J. E. Warren, M. Xu and N. B. McKeown, *Angew. Chem., Int. Ed.*, 2009, **48**, 3273–3277.

33 J. D. Evans, D. M. Huang, M. Haranczyk, A. W. Thornton, C. J. Sumby and C. J. Doonan, *CrystEngComm*, 2016, **18**, 4133–4141.

34 M. Miklitz, S. Jiang, R. Clowes, M. E. Briggs, A. I. Cooper and K. E. Jelfs, *J. Phys. Chem. C*, 2017, **121**, 15211–15222.

35 A. Li, R. Bueno-Perez and D. Fairen-Jimenez, *Chem. Sci.*, 2022, **13**, 13507–13523.

36 S. Grimme, C. Bannwarth and P. Shushkov, *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.

37 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.

38 M. Bursch, H. Neugebauer and S. Grimme, *Angew. Chem.*, 2019, **131**, 11195–11204.

39 S. Schmitz, J. Seibert, K. Ostermeir, A. Hansen, A. H. Göller and S. Grimme, *J. Phys. Chem. B*, 2020, **124**, 3636–3646.

40 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

41 J. VandeVondele and J. Hutter, *J. Chem. Phys.*, 2007, **127**, 114105.

42 S. Grimme, S. Ehrlich and L. Goerigk, *J. Chem. Phys.*, 2011, **32**, 1456–1465.

43 M. Liu, M. A. Little, K. E. Jelfs, J. T. A. Jones, M. Schmidtmann, S. Y. Chong, T. Hasell and A. I. Cooper, *J. Am. Chem. Soc.*, 2014, **136**, 7583–7586.

44 *Materials Studio (version 2020) BIOVIA*, Dassault Systèmes, San Diego, 2020.

45 D. M. D'Alessandro, B. Smit and J. R. Long, *Angew. Chem., Int. Ed.*, 2010, **49**, 6058–6082.

46 P. G. Boyd, A. Chidambaram, E. García-Díez, C. P. Ireland, T. D. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S. M. Moosavi, M. M. Maroto-Valer, J. A. Reimer, J. A. R. Navarro, T. K. Woo, S. Garcia, K. C. Stylianou and B. Smit, *Nature*, 2019, **576**, 253–256.

47 P. Z. Moghadam, Y. G. Chung and R. Q. Snurr, *Nat. Energy*, 2024, **9**, 121–133.

48 M. Mastalerz and I. M. Oppel, *Angew. Chem., Int. Ed.*, 2012, **51**, 5252–5255.

49 Z. Wang, N. Sikdar, S.-Q. Wang, X. Li, M. Yu, X.-H. Bu, Z. Chang, X. Zou, Y. Chen, P. Cheng, K. Yu, M. J. Zaworotko and Z. Zhang, *J. Am. Chem. Soc.*, 2019, **141**, 9408–9414.

50 K. Su, W. Wang, S. Du, C. Ji and D. Yuan, *Nat. Commun.*, 2021, **12**, 3703.

51 K. Tian, S. M. Elbert, X. Hu, T. Kirschbaum, W. Zhang, F. Rominger, R. R. Schröder and M. Mastalerz, *Adv. Mater.*, 2022, **34**, 2202290.

52 A. Fauzi Ismail, K. Khulbe and T. Matsuura, *Gas Separation Membranes: Polymeric and Inorganic*, Springer, 2015.

53 S. T. Schneebeli, M. Frasconi, Z. Liu, Y. Wu, D. M. Gardner, N. L. Strutt, C. Cheng, R. Carmieli, M. R. Wasielewski and J. F. Stoddart, *Angew. Chem., Int. Ed.*, 2013, **52**, 13100–13104.

54 Y. Beldjoudi, A. Narayanan, I. Roy, T. J. Pearson, M. M. Cetin, M. T. Nguyen, M. D. Krzyaniak, F. M. Alsubaie, M. R. Wasielewski, S. I. Stupp and J. F. Stoddart, *J. Am. Chem. Soc.*, 2019, **141**, 17783–17795.

55 B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshaye, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. Van Der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W.-z. Yu and T. Frauenheim, *J. Chem. Phys.*, 2020, **152**, 124101.

56 T. A. Manz and D. S. Sholl, *J. Chem. Theory Comput.*, 2010, **6**, 2455–2468.

57 T. A. Manz and N. G. Limas, *RSC Adv.*, 2016, **6**, 47771–47801.

58 T. A. Manz and D. S. Sholl, *J. Chem. Theory Comput.*, 2010, **6**, 2455–2468.

59 T. A. Manz and D. S. Sholl, *J. Chem. Theory Comput.*, 2012, **8**, 2844–2867.

60 J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing and J. Hutter, *Comput. Phys. Commun.*, 2005, **167**, 103–128.

61 S. Goedecker, M. Teter and J. Hutter, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 1703–1710.

62 T. Lu and F. Chen, *J. Comput. Chem.*, 2012, **33**, 580–592.

63 T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, *Microporous Mesoporous Mater.*, 2012, **149**, 134–141.

64 D. Dubbeldam, S. Calero, D. E. Ellis and R. Q. Snurr, *Mol. Simul.*, 2016, **42**, 81–101.

65 S. L. Mayo, B. D. Olafson and W. A. Goddard, *J. Phys. Chem.*, 1990, **94**, 8897–8909.

66 J. J. Potoff and J. I. Siepmann, *AIChE J.*, 2001, **47**, 1676–1682.

67 M. G. Martin and J. I. Siepmann, *J. Phys. Chem. B*, 1998, **102**, 2569–2577.